Supplementary Material

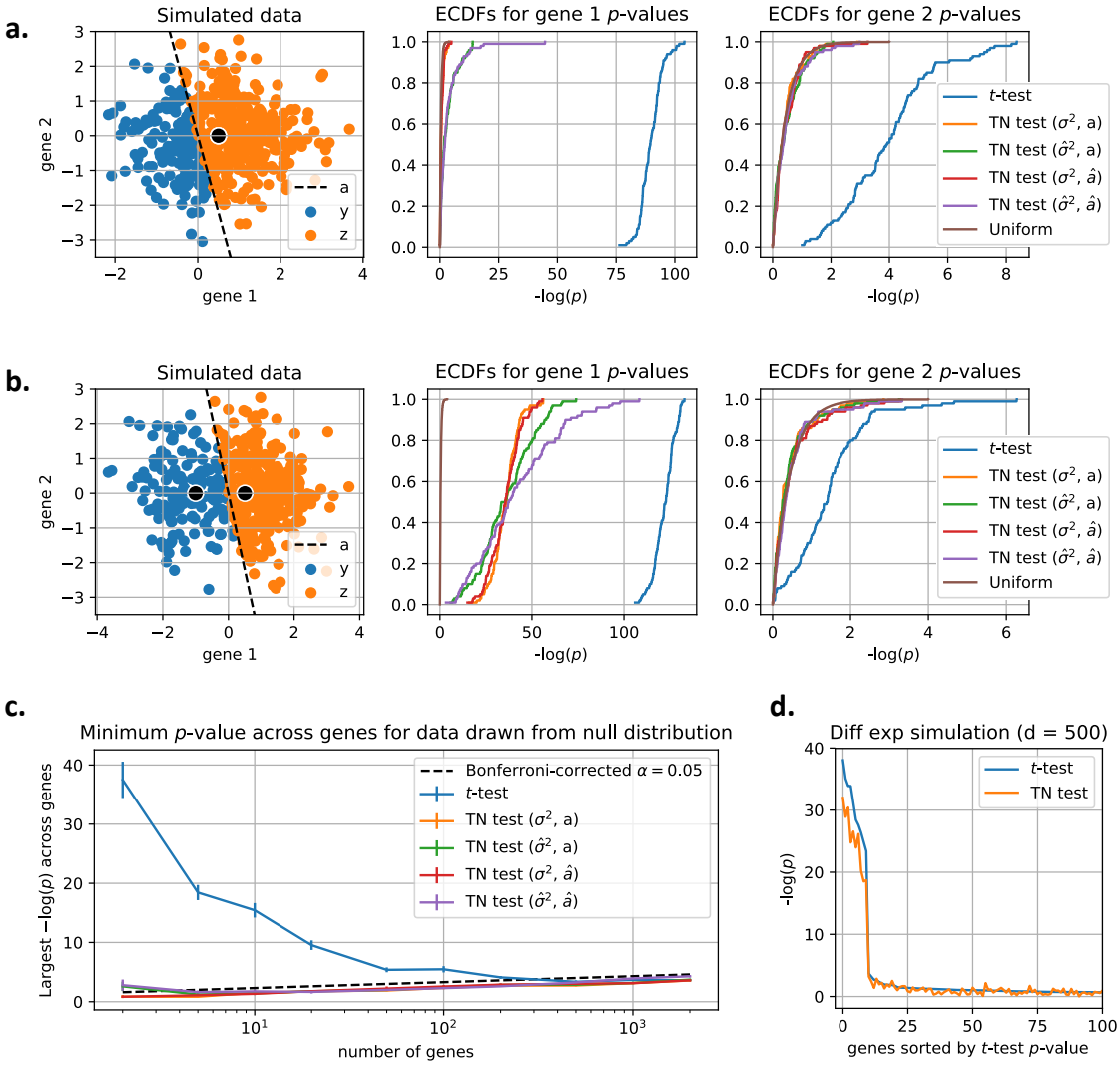

Figure 1: **Results on simulated data drawn from truncated normal distributions, Related to STAR Methods. a.** 500 samples are drawn from the same distribution, and genes 1 and 2 are drawn from $\mathcal{N}(0.5, 1)$ and $\mathcal{N}(0, 1)$, respectively. The clustering step splits the dataset into groups of 156 and 344 samples, and a exactly captures the clustering rule. We see that although neither gene is differentially expressed in the underlying distribution, the t-test consistently returns small p-values across 100 simulation runs. We present four versions of the TN test, all of which significantly correct for the clustering step. $\hat{\sigma}^2$ indicates that the variance was unknown and therefore estimated from the data. $\hat{a}$ indicates that the hyperplane was estimated from a held-out 10% of the samples using an SVM. **b.** The experiment from **a** is repeated except gene 1 is drawn from a $\mathcal{N}(-1, 0)$ distribution instead for one of the clusters. The number of samples in each group and the separating hyperplane remain the same. **c.** We explore how the minimum p-value across genes changes with d, the number of genes. For a particular number of genes, 200 samples are drawn from a $\mathcal{N}(0, I)$ distribution, and a is chosen randomly. This simulation is repeated 10 times for each value of d. α indicates the chosen level of significance. **d.** For $d = 500$, we run a 200-sample simulation experiment (100 in each cluster) where 10 genes are differentially expressed. 10 values of μ_L were set to -1, and the corresponding entries in μ_R were set to 1. All other entires of μ_L, μ_R were set to 0, and $\sigma^2 = 1$.

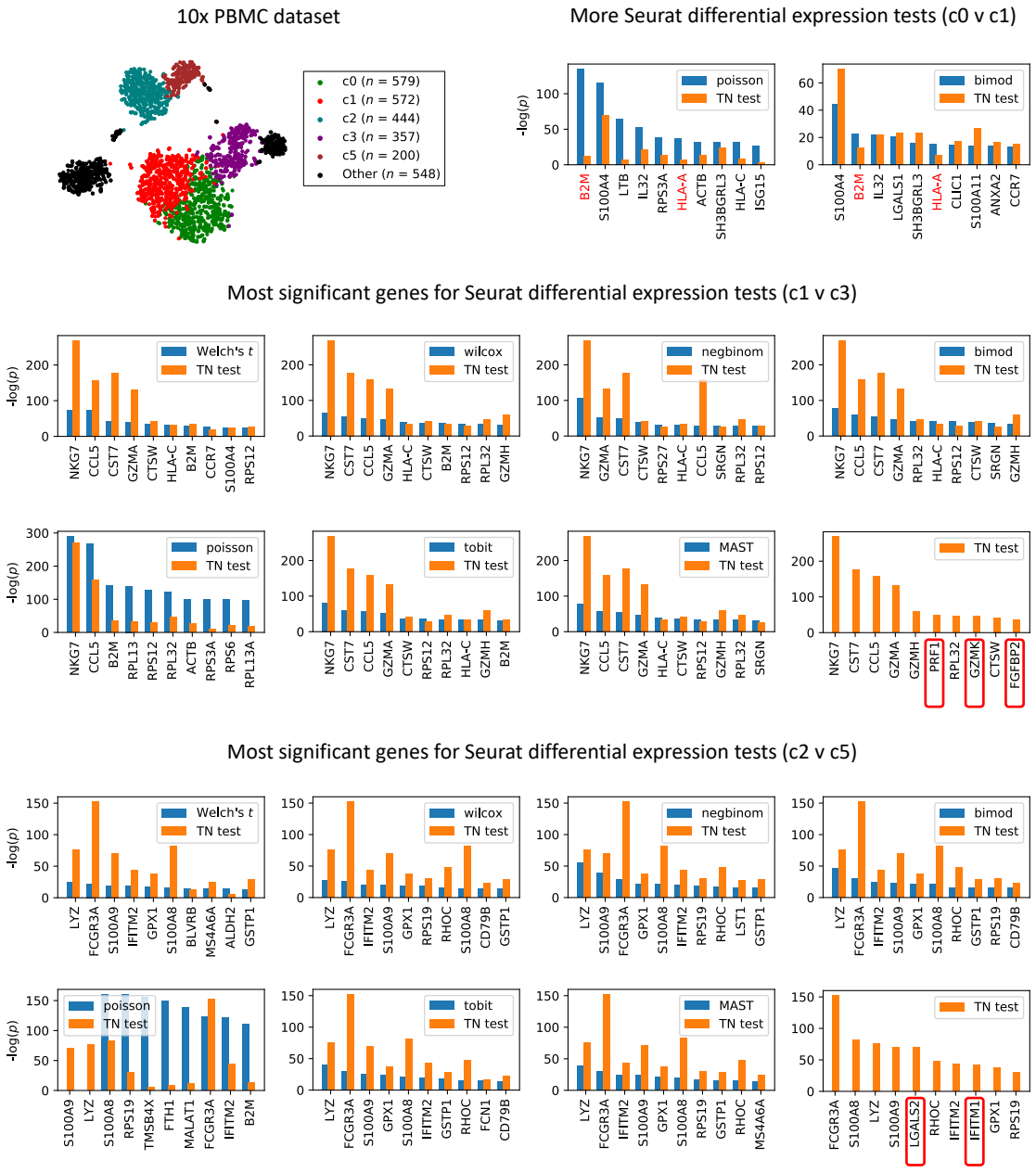

Figure 2: **Comparison of differential expression tests on PBMC dataset continued, Related to Figure 2a.** The comparison performed in Figure 2a is repeated for other Seurat differential expression methods and for clusters 1 v 3 and 2 v 5. A missing bar indicates a p value of 0 due to numerical precision limitations. TN test genes boxed in red were missed by the other tests.

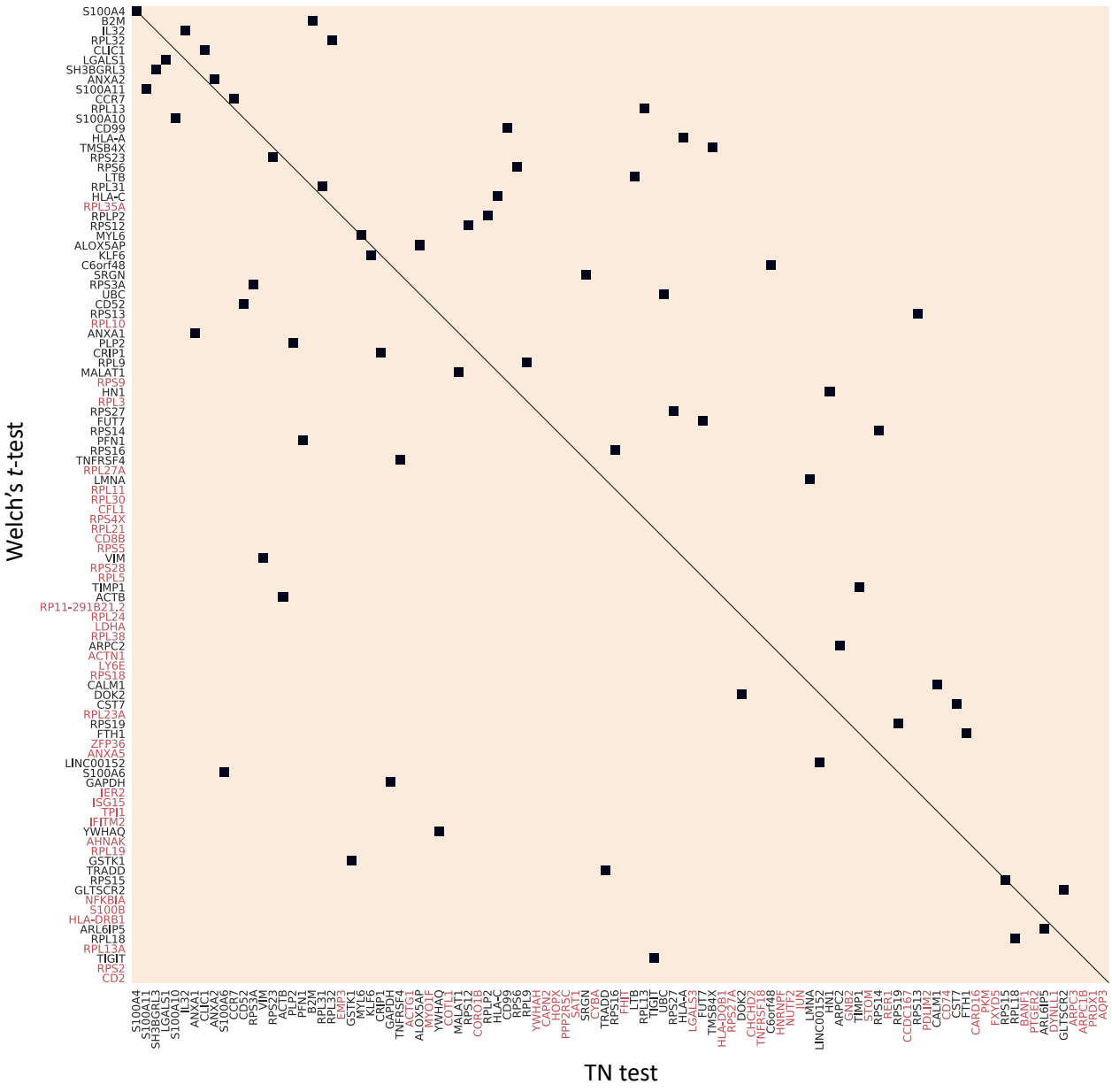

Figure 3: **Gene set comparison using permutation matrix, Related to Figure 2a.** The top 100 genes selected by a Welch's t-test compared to the top 100 genes selected by the TN test when performing differential expression analysis on clusters 0 and 1 in Fig. 2a. Genes written in red text were not selected by the other test. 64 genes appear in both gene sets.

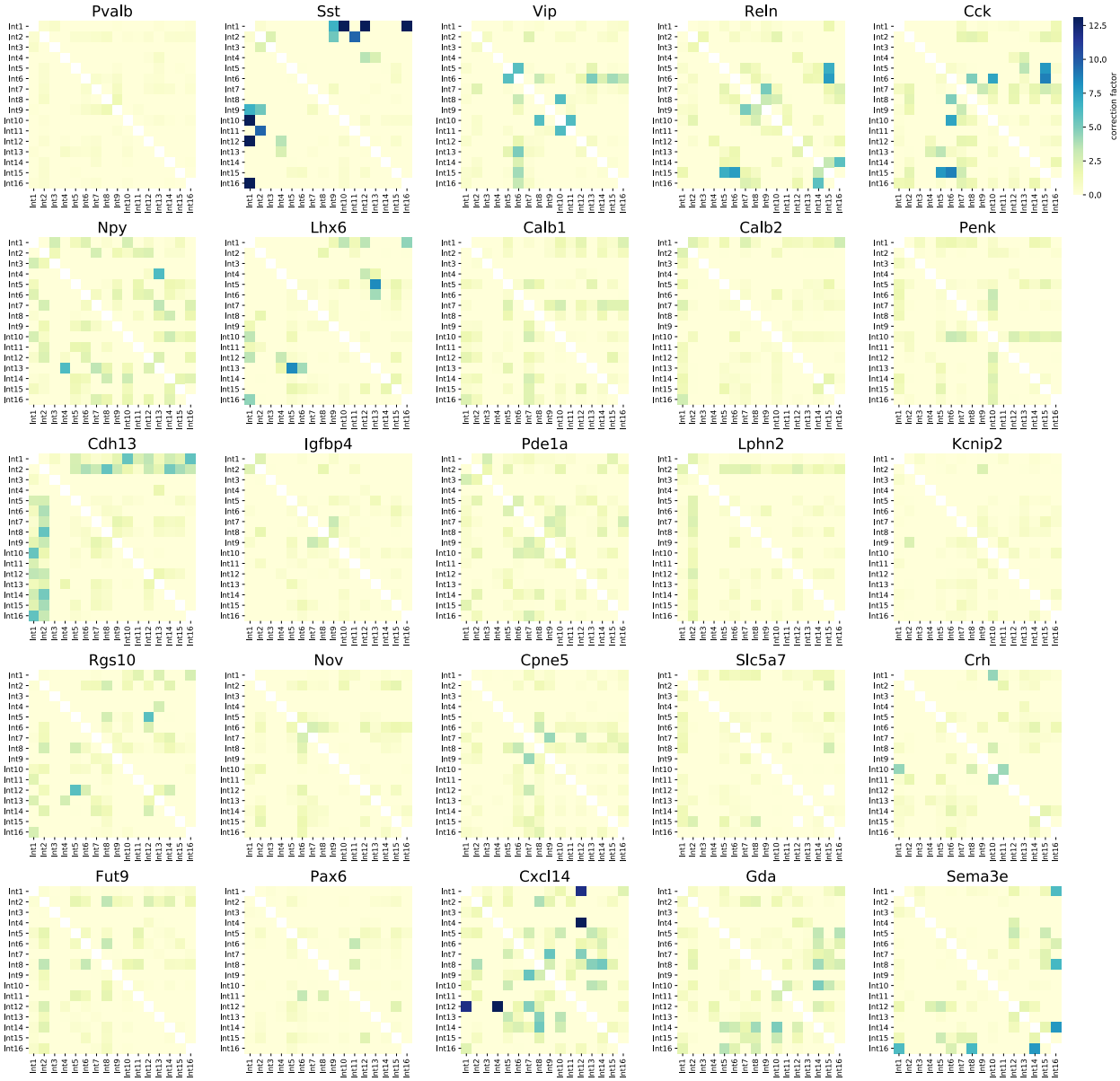

Figure 4: **TN test correction for mouse brain cell dataset interneuron genes, Related to Figure 3a.** The 16 interneuron subclasses reported for the mouse brain cell dataset Zeisel et al. 2015 are re-compared pairwise using each of the 26 genes discussed by the authors. For each gene and pair of subclasses, the correction factor represents the -log of the ratio of the t-test p-value to the TN test p-value. We only consider comparisons where the hyperplane fit the data relatively well (58.3% of comparisons).